Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

Automatic Video Captioning in TRECVid's Video-to-Text Task

Alan F. Smeaton
Dublin City University

A world leading SFI Research Centre

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
EUROPEAN UNION
European Regional Development Fund
DCU
NUI Galway OÉ Gaillimh
UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh
UCD DUBLIN
Science Foundation Ireland sfi For what's next

Motivation

Exponential increase of generated multimedia content

https://bit.ly/2OZ1jx3

Motivation

Insight
Centre for Data Analytics

...keeping a record of memorable personal moments...

https://bit.ly/2OZ1jx3

Pope Francis @ Philippines, 2015 (Source: AP Photo/Bullit Marquez)

Motivation

Insight
Centre for Data Analytics

...keeping a record of memorable moments https://bit.ly/2OZ1jx3

Pope Francis @ Ecuador, 2015 (Source: AP)

Motivation

Insight
Centre for Data Analytics

…(or not).

https://bit.ly/2OZ1jx3

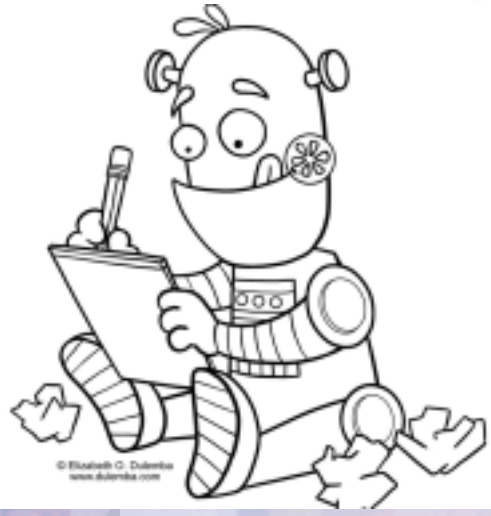
Pope Francis @ USA, 2015

Manual vs Automatic Annotation

Annotation is the process of generating high level metadata (semantic).

How to generate semantic metadata ?

Manual Annotation



Automatic Annotation

Manual Annotation

Insight
Centre for Data Analytics

Problem: Manual Annotation is tedious. https://bit.ly/2OZ1jx3

Manual Annotation

Insight
Centre for Data Analytics

Annotation can be split and assigned to the **crowd** as...

Concepts in image annotation

- Well developed in earlier TRECVid task and in ImageNet, and elsewhere
- Google+ photos now uses computer vision and machine learning to identify objects and settings in your uploaded snapshots
- Google have learned 000's (visual) concepts and apply them to personal (you and your friends') photos
- Others followed .. Apple have it on your iPhone !

NY Times, Nov 2014, premature…

Insight
Centre for Data Analytics

Researchers Announce Advance in Image-Recognition Software

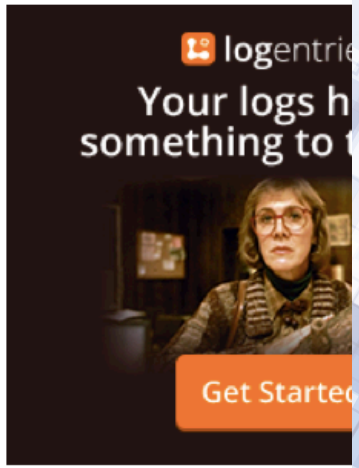
By JOHN MARKOFF NOV. 17, 2014

MOUNTAIN VIEW, Calif. — Two groups of scientists, working independently, have created artificial intelligence software capable of recognizing and describing the content of photographs and videos with far greater accuracy than ever before, sometimes even mimicking human levels of understanding.

Until now, so-called computer vision has largely been limited to recognizing individual objects. The new software, described on Monday by researchers at Google and at Stanford University, teaches itself to identify entire scenes: a group of young men playing Frisbee, for example, or a herd of elephants marching on a grassy plain.

The software then writes a caption in English describing the picture. Compared with human observations, the researchers found, the computer-written descriptions are surprisingly accurate.

The advances may make it possible to better catalog and search for the billions of images and hours of video available online, which are often poorly described and archived. At the moment, search engines like Google rely largely on written language accompanying an image or video to ascertain what it contains.



RELATED COVERAGE

Computer Eyesight Accurate AUG. 18, 20

Image captions (not tags)

A group of young men playing a game of frisbee

Captioning was done by sub-frame tagging…

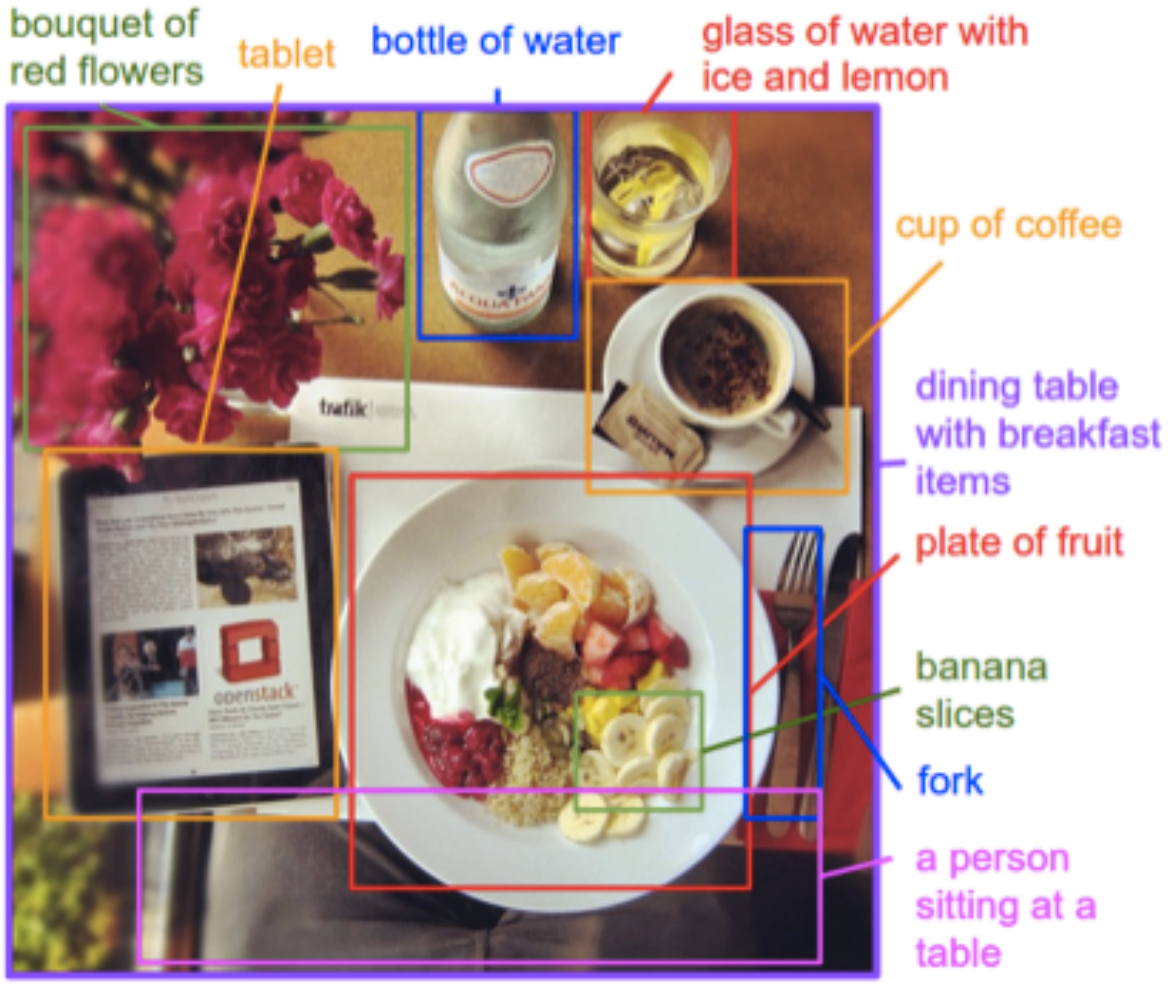


Figure 1. Our model generates free-form natural language descriptions of image regions.

Figure 5. Example alignments predicted by our model. For every test image above, we retrieve the most compatible test sentence and visualize the highest-scoring region for each word (before MRF smoothing described in Section 3.1.4) and the associated scores ($v_i^T s_t$). We hide the alignments of low-scoring words to reduce clutter. We assign each region an arbitrary color.

Now, everybody is using deep learning, for everything

Sometimes it works OK, sometimes its really good, it's the dominant approach in image, and video, captioning

TRECVid is …

- A global benchmark, running annually since 2001
- Hosted and run by US National Institute of Standards and Technology
- Founded and is co-led by myself and Wessel Kraaij (TNO Netherlands)
- Addresses content-based tasks on video …
  - Shot boundary detection, video summarisation, semantic concept detection, ad hoc search, known item search, copy detection, surveillance events, multimedia event detection, video hyperlinking, localisation search … and … video-to-text
- Has open participation and global engagement, with +2,000 researchers directly involved since it started 17 years ago
- VTT – very recent work

VTT Goals and Motivations

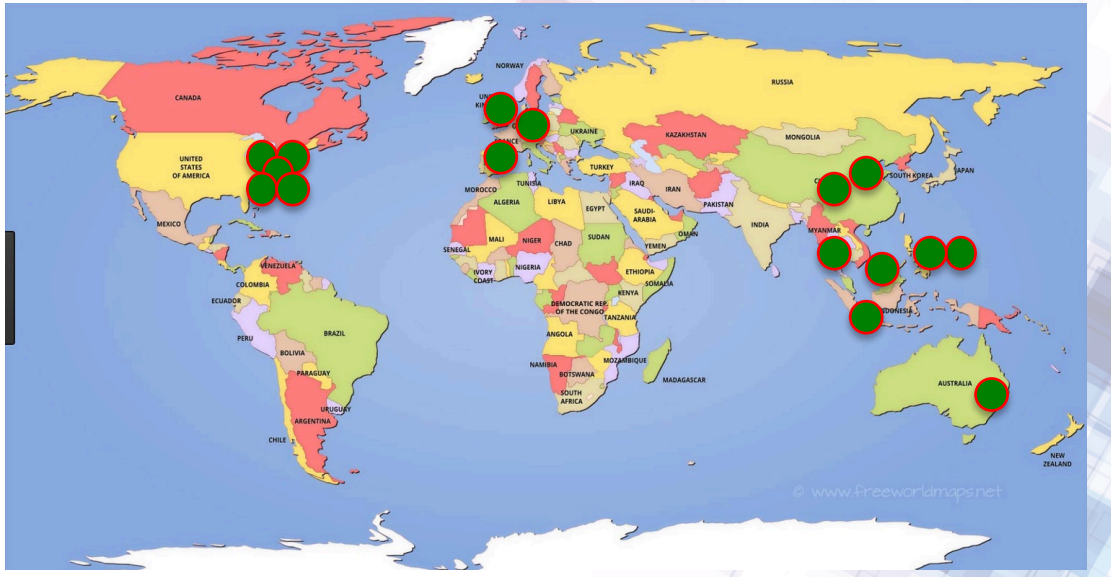
- ✓ Measure how well an automatic system can describe a video in natural language.
- ✓ Transfer successful image captioning technology to the video domain.

Real world Applications

- ✓ Video summarization
- ✓ Supporting search and browsing
- ✓ Accessibility - video description to the blind

Following a pilot in 2016, who took part in VTT in 2017 ?

- University of Amsterdam
- Carnegie Mellon University
- National University of Singapore
- City University of Hong Kong
- City College of New York
- University of Technology, Sydney
- Shandong University, China
- Tianjun University, China
- Renmin University, China
- Korea University
- UPC Barcelona
- National Institute of Informatics (Japan)
- Hitachi (Japan)
- Two US-based R&D companies, Aréte and Etter

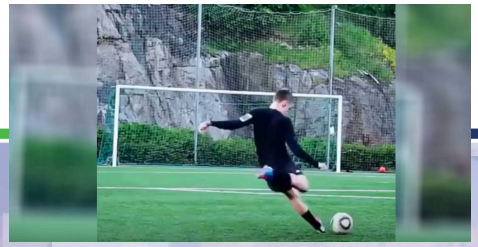


… and Dublin City University

Video Dataset

- 50k+ Twitter Vine video URLs, 6s max
- A subset of 1,880 randomly selected, **manually captioned**
  1. Some complex scenes contain a lot of information to describe.
  2. Assessors interpret scenes according to cultural or pop cultural references, not universally recognized.
  3. There are some similar videos, resulting in similar descriptions
     - **Visual similarity** using CUHK Bag of Visual Words to cluster and remove near duplicates and visually similar (e.g. soccer games)
     - **Description similarity** was detected using caption clustering and manual removal

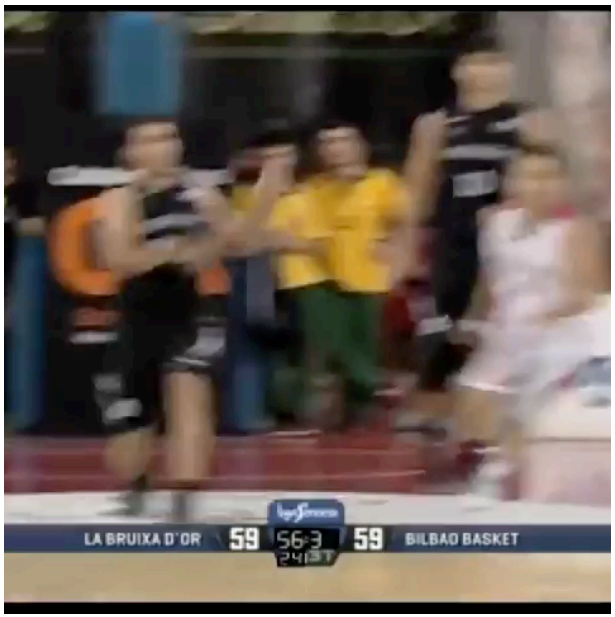





Sample Manual Captions

1. Many people hold long trampoline and person does double somersault.
2. A group of men hoist a man into the air and he does a flip.
3. Group of young men holding a portable trampoline/mat and when they raise it man on top of trampoline flips and somersaults into the air and lands on his feet.
4. Man thrown in air, manages at least five head over heels in high somersault.
5. One trampoline athlete demonstrates perfectly.

Sample Manual Captions



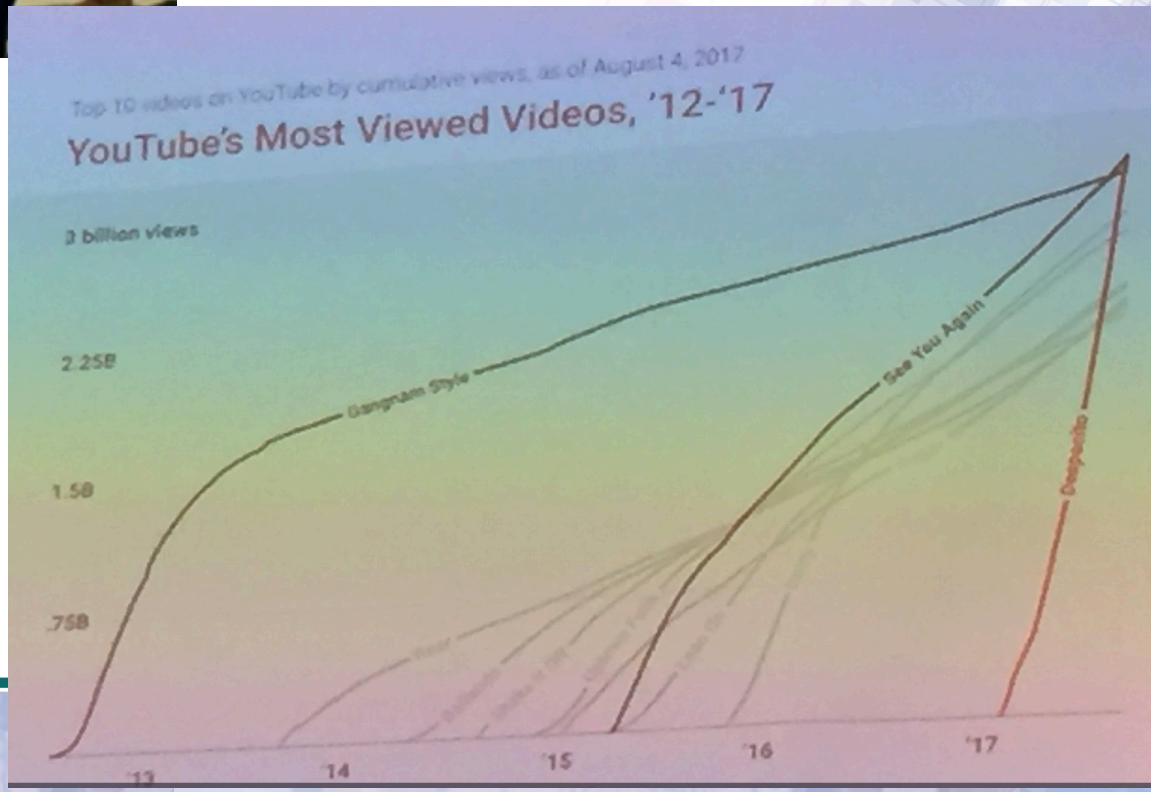
1. Basketball player misses shot, goes out of bounds, and teammate makes basket and physically hangs onto basket for a time.
2. A basketball player hangs on the basket, at basketball play.
3. A basketball player is barreling towards the basket when he is sideswiped by and opponent looses control of the ball; his teammate recovers the basketball, scores for two points and swings from the basketball rim.
4. A player scored a point in a basketball game.
5. Basketball game in progress; black jersey player makes basket and hangs on rim.

Annotation Process – Observations

1. Some complex scenes contain a lot of information to describe.
2. **Assessors interpret scenes according to cultural or pop cultural references, not universally recognized.**
3. Specifying the time of the day was often not possible for indoor videos.
4. There may be some similar videos, resulting in similar descriptions. This was minimized by redundancy removal.

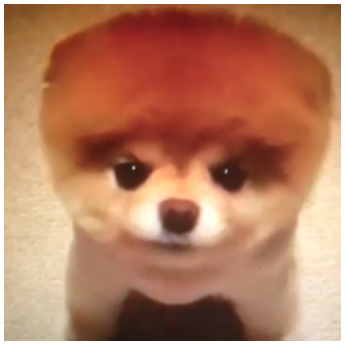
Insight
Centre for Data Analytics

+5,484,591,787 views !



Description Generation

Given a video …



Generate a textual description

Who ? What ? Where ? When ?

"a dog is licking its nose"

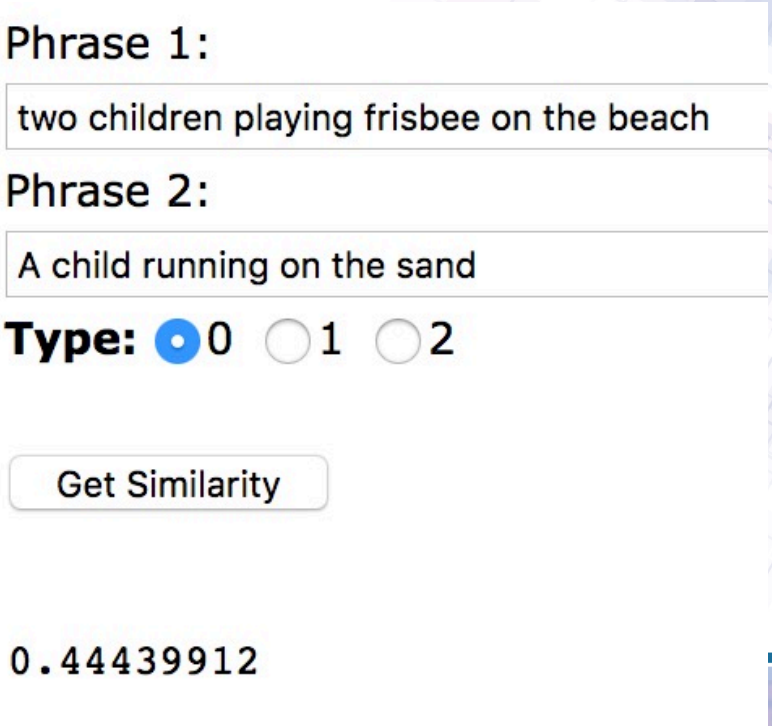
Metrics

- Conventionally popular MT measures : BLEU, METEOR, CIDEr
- Each site asked to nominate one run as "primary"

Insight
Centre for Data Analytics

Metrics

- **Semantic Text Similarity (STS)** – based on distributional similarity and Latent Semantic Analysis (LSA) … complemented with semantic relations extracted from WordNet



Direct Assessment (DCU)

- Brings human (AMT) into the evaluation by crowdsourcing how well a caption describes a video … rate a caption [0..100]
- Automatically degrade the quality of some manual captions to rate the quality of the assessors – distinguish genuine from those gaming the system
- A variation on what is used in the main benchmark in MT, the *Workshop on Statistical Machine Translation* (WMT)
- Re-ran this on VTT 2016 submissions, twice, with 0.99 correlation on scores and rankings, showing consistency

Evaluation of automatic video captioning using direct assessment

Yvette Graham[1], George Awad[2,3], Alan Smeaton[4]*

1 ADAPT Centre for Digital Content Technology, Dublin City University, Glasnevin, Dublin 9, Ireland, **2** National Institute of Standards and Technology, Gaithersburg, MD, United States of America, **3** Dakota Consulting, Inc., Silver Spring, MD, United States of America, **4** Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland

* alan.smeaton@dcu.ie

PLOS ONE

Received: August 23, 2017

Accepted: August 9, 2018

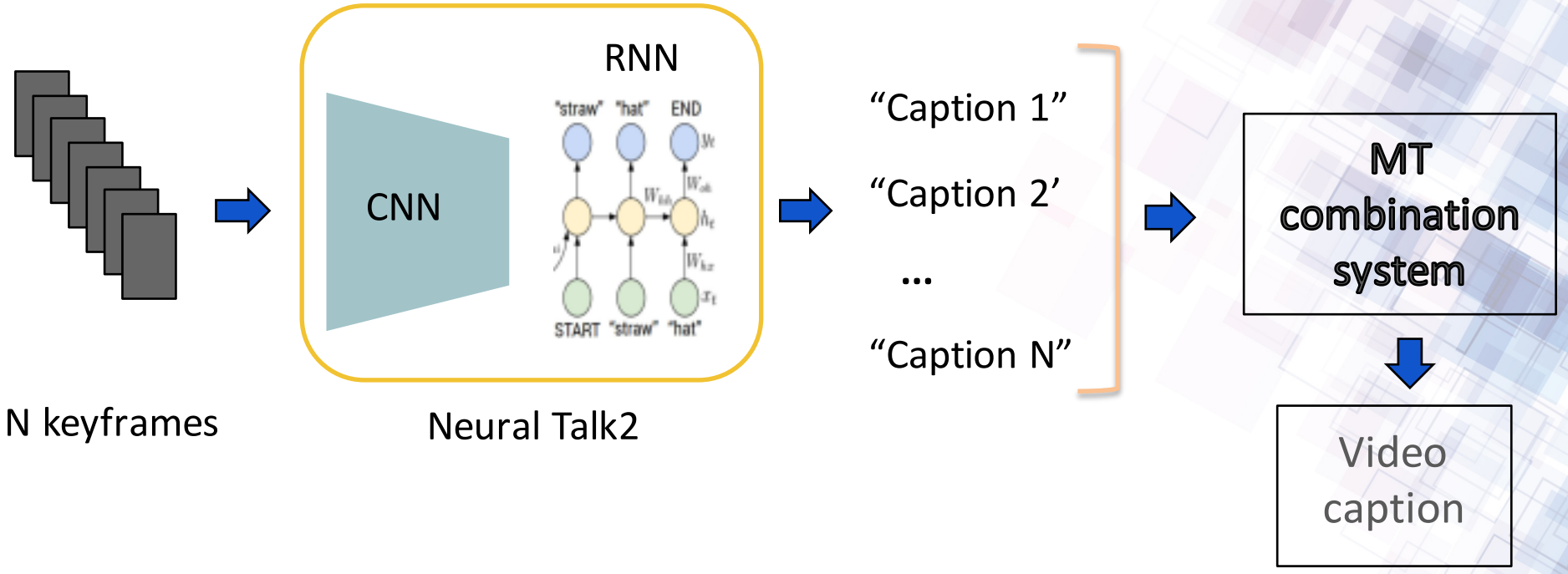
Published: September 4, 2018

An example from run submissions – unique examples

- A woman holding a microphone
- A woman is dancing
- A woman wearing a hat is singing into a microphone
- A woman sings on a stage
- A girl is singing on a stage
- A woman is singing a song
- A woman is singing a song on stage in a beauty salon
- A woman is talking to a man

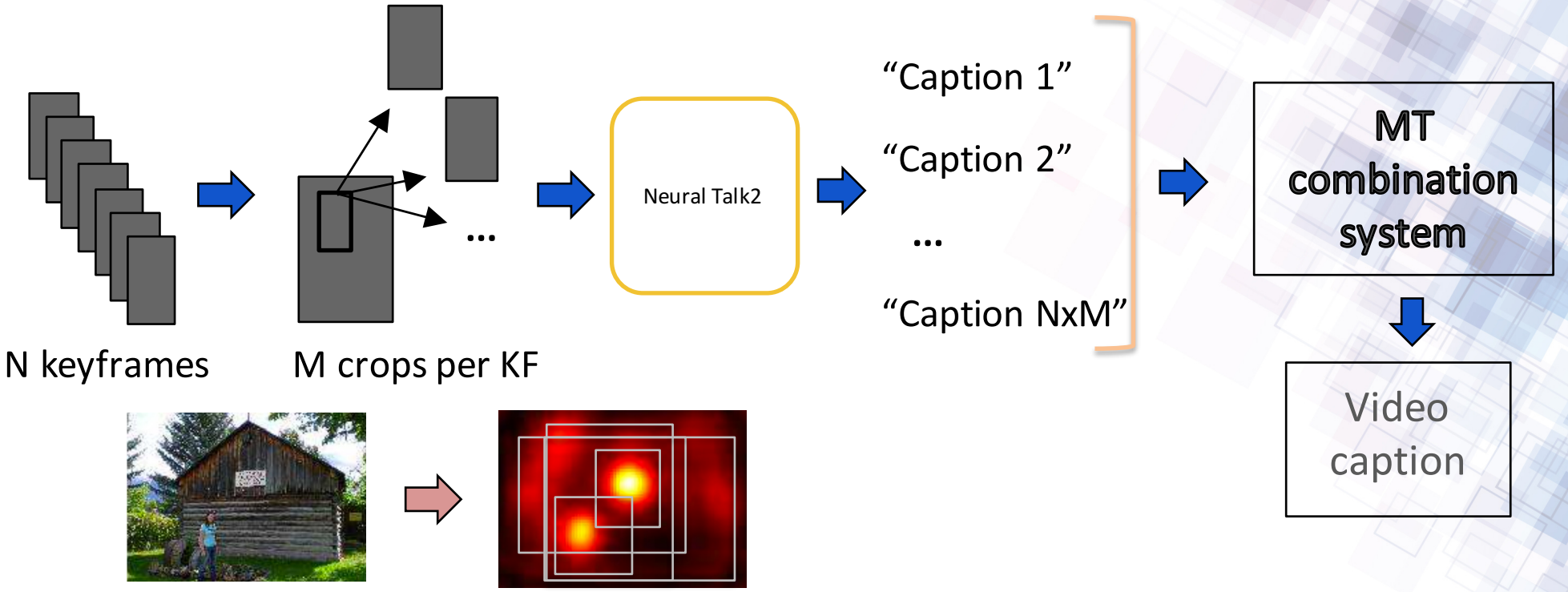
As an example … DCU 1/3

- One caption generated for each keyframe.



DCU 2/3

- M crops based on spatial salience extracted for each keyframe.
- One caption is generated for each crop.

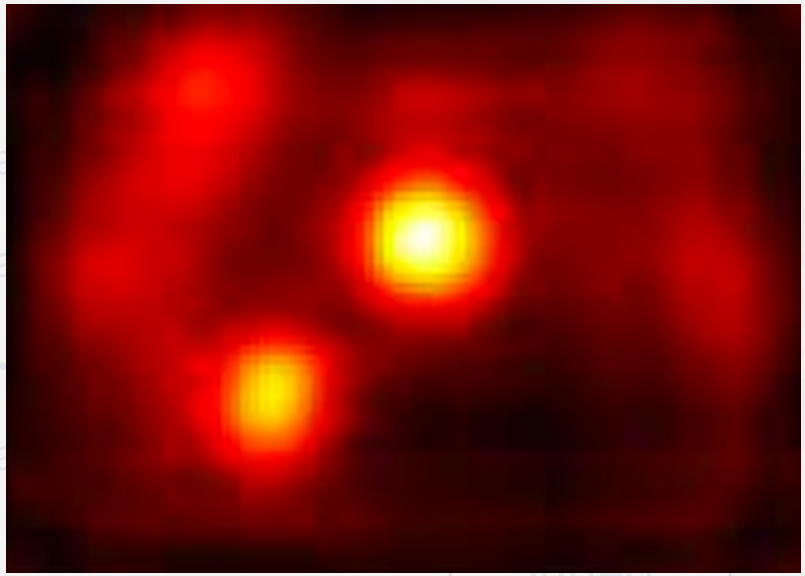


DC

- xtracted for each keyframe.
- rop.

DC

- xtracted for each keyframe.
- rop.



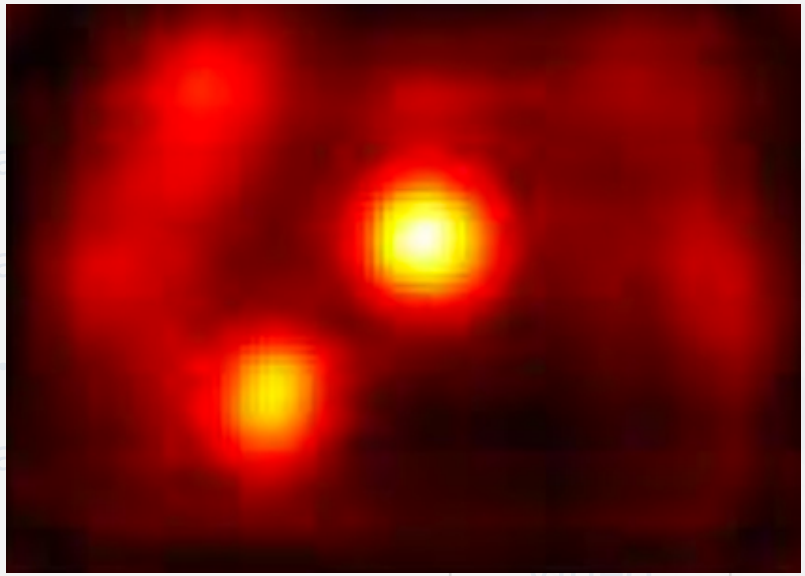
"Ca
"Ca
"Ca

N keyframes M crops per KF

Video caption

DC

- xtracted for each keyframe.
- rop.



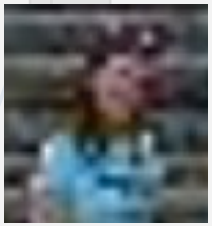
N keyframes

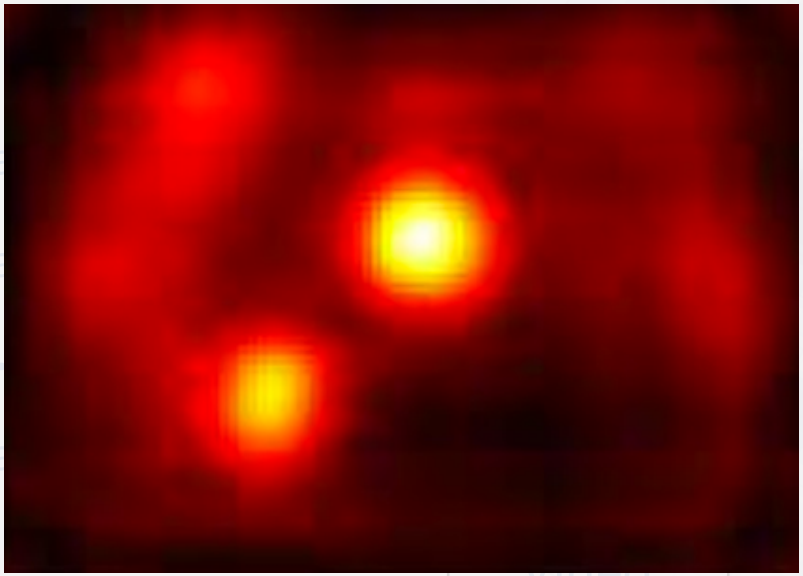
M crops per KF

Neural Talk2

"Ca
"Ca
"Ca

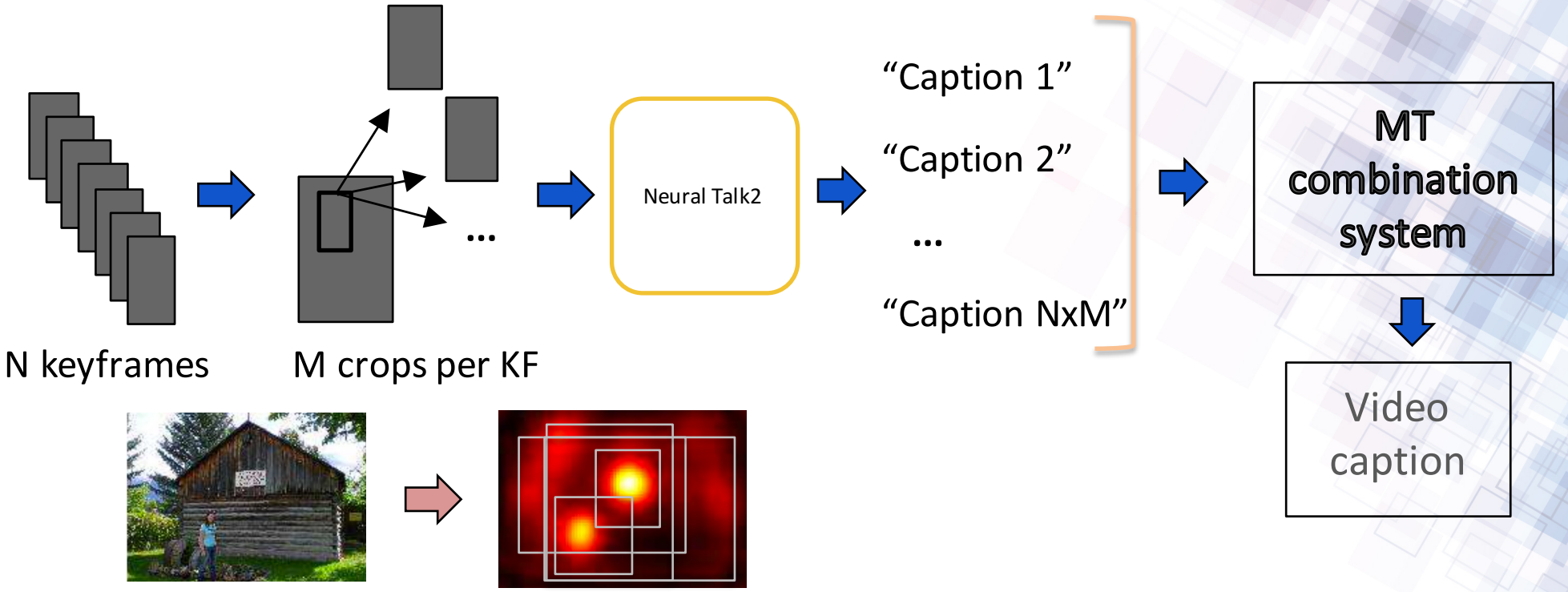
Video caption





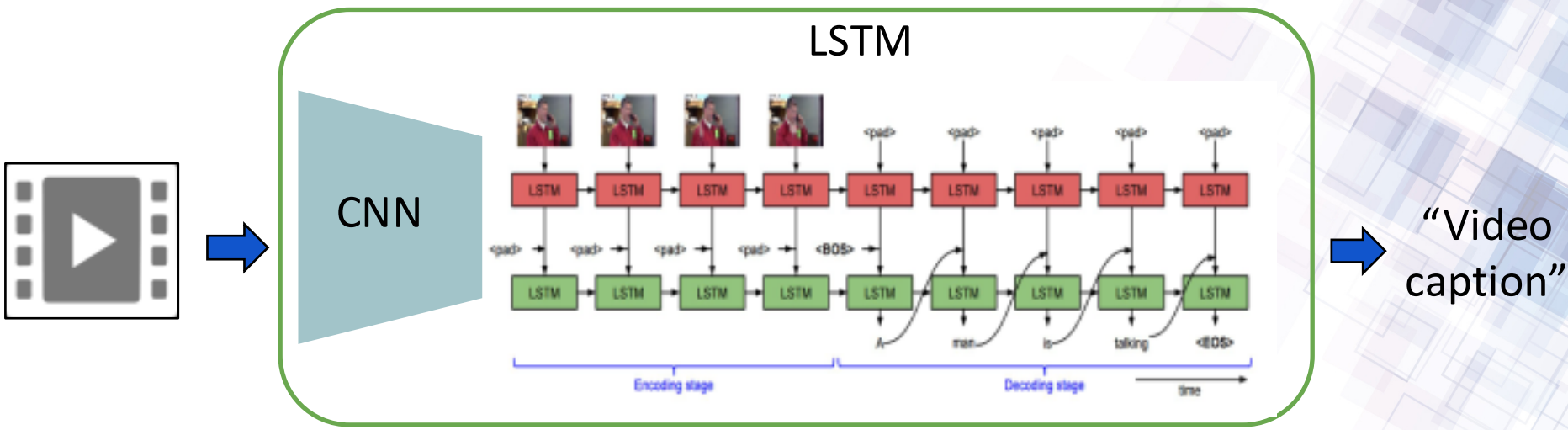
DCU 2/3

- M crops based on spatial salience extracted for each keyframe.
- One caption is generated for each crop.



DCU 3/3

- Video features generated with a CNN, passed to a 2x LSTM stack
- LSTM's encode the features and decode into natural language descriptions



Sequence to Sequence - Video to Text (S2VT)

Some Insight - ADAPT automatic captions …

#990
a baseball player holding a bat on a field

#1599
a white cat sitting on top of a table

#603
a green truck is parked on a street

#1695
a person riding a bike down a street

TRECVid VTT Results …

Systems Rankings for each Metric

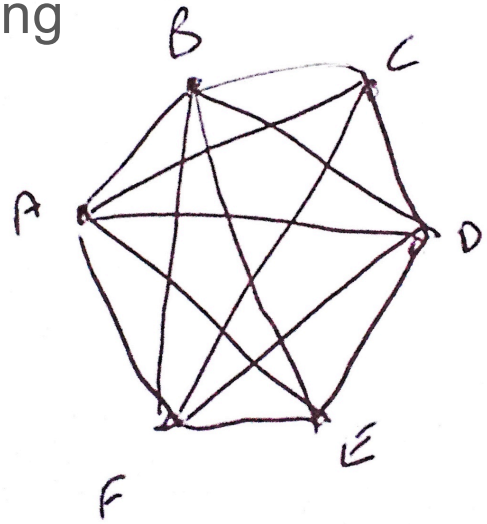
Insight
Centre for Data Analytics

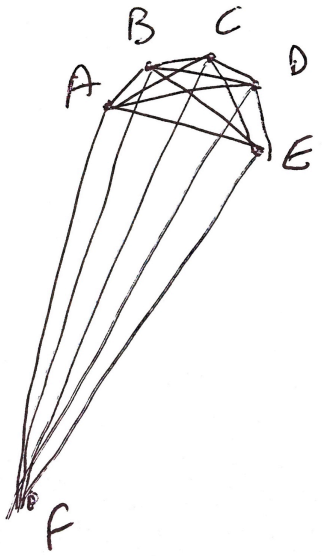
CIDEr	METEOR	BLEU	STS
RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU
mediamil	mediamil	mediamil	INF
INF	INF	TJU	mediamil
TJU	DCU	UTS_CAI	NII_Hitachi_UIT
UTS_CAI	TJU	INF	TJU
VIREO	VIREO	DCU	UTS_CAI
NII_Hitachi_UIT	UTS_CAI	VIREO	VIREO
ARETE	KU_ISPL	NII_Hitachi_UIT	CCNY
DCU	SDNU_MMSys	SDNU_MMSys	SDNU_MMSys
SDNU_MMSys	NII_Hitachi_UIT	CCNY	KU_ISPL
CCNY	ARETE	ARETE	DCU
KU_ISPL	CCNY	KU_ISPL	ARETE
UPCer	UPCer	UPCer	UPCer

Insight
Centre for Data Analytics

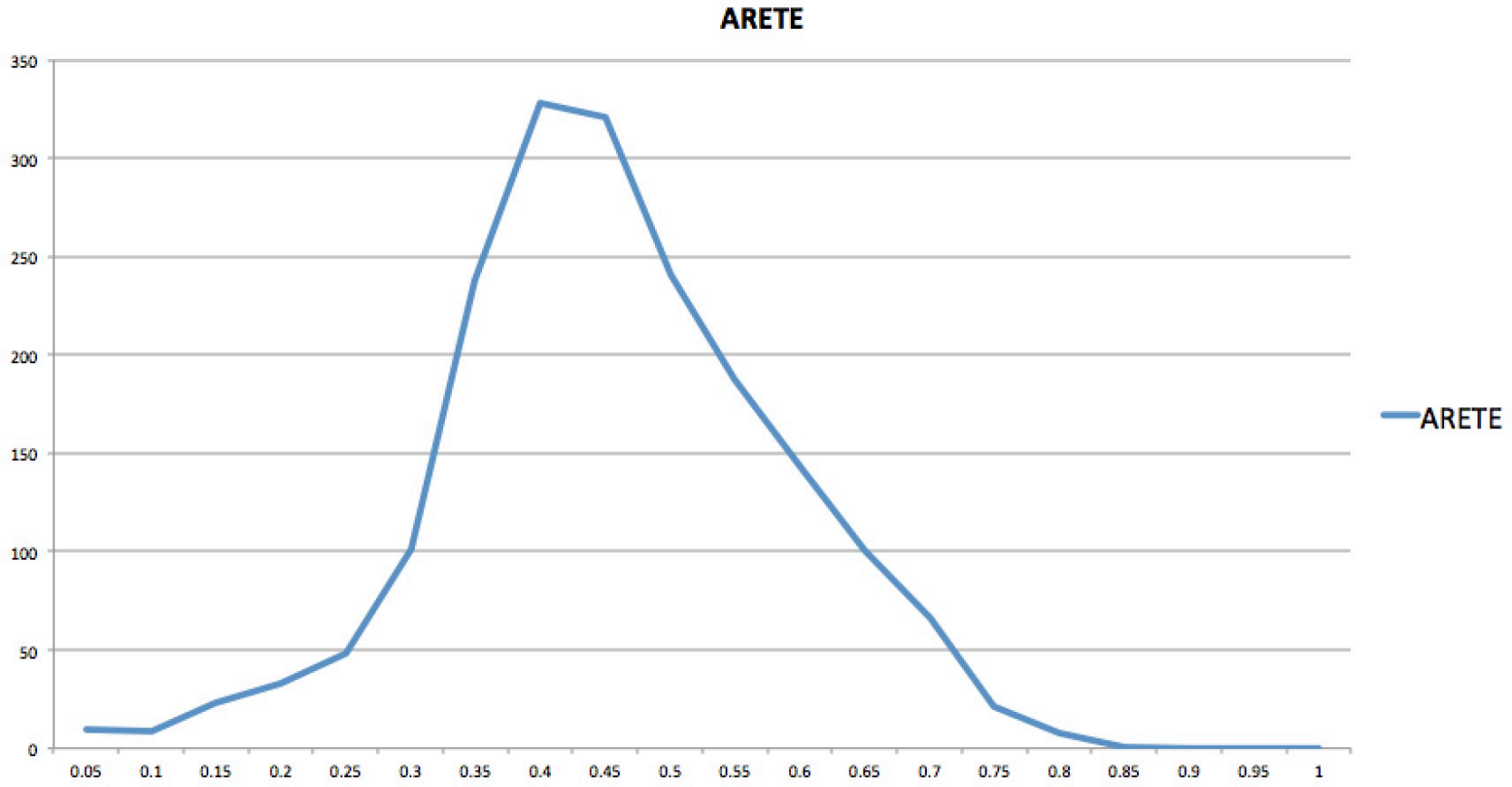
STS Results - Analysis

- METEOR / BLEU / CIDER / STS of runs vs. manual is meaningless … manual is a **reference**, not a groundtruth
- So we measured, for each video, pairwise similarity among all submissions (primary run only) for 13 systems + 1 manual (171,080 pairwise comparisons - thanks to UMBC)
- Ideally all systems very similar but the more “outlier-ish” a system, across all 1,880 videos (lower averaged STS value), says something



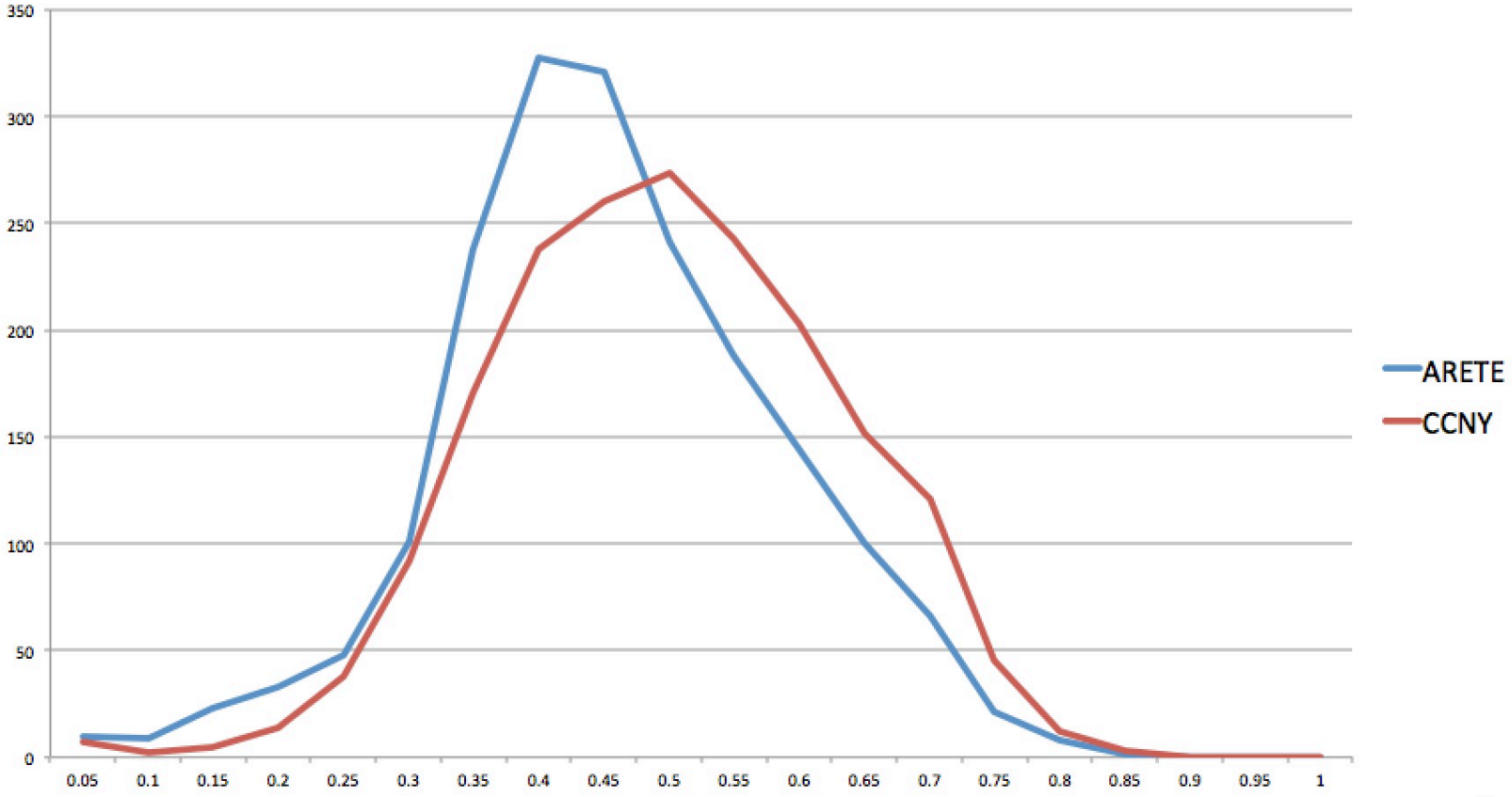


Take ARETE … for each 1,880 videos, compute STS vs. each other system (+human), value into 1 of 20 buckets



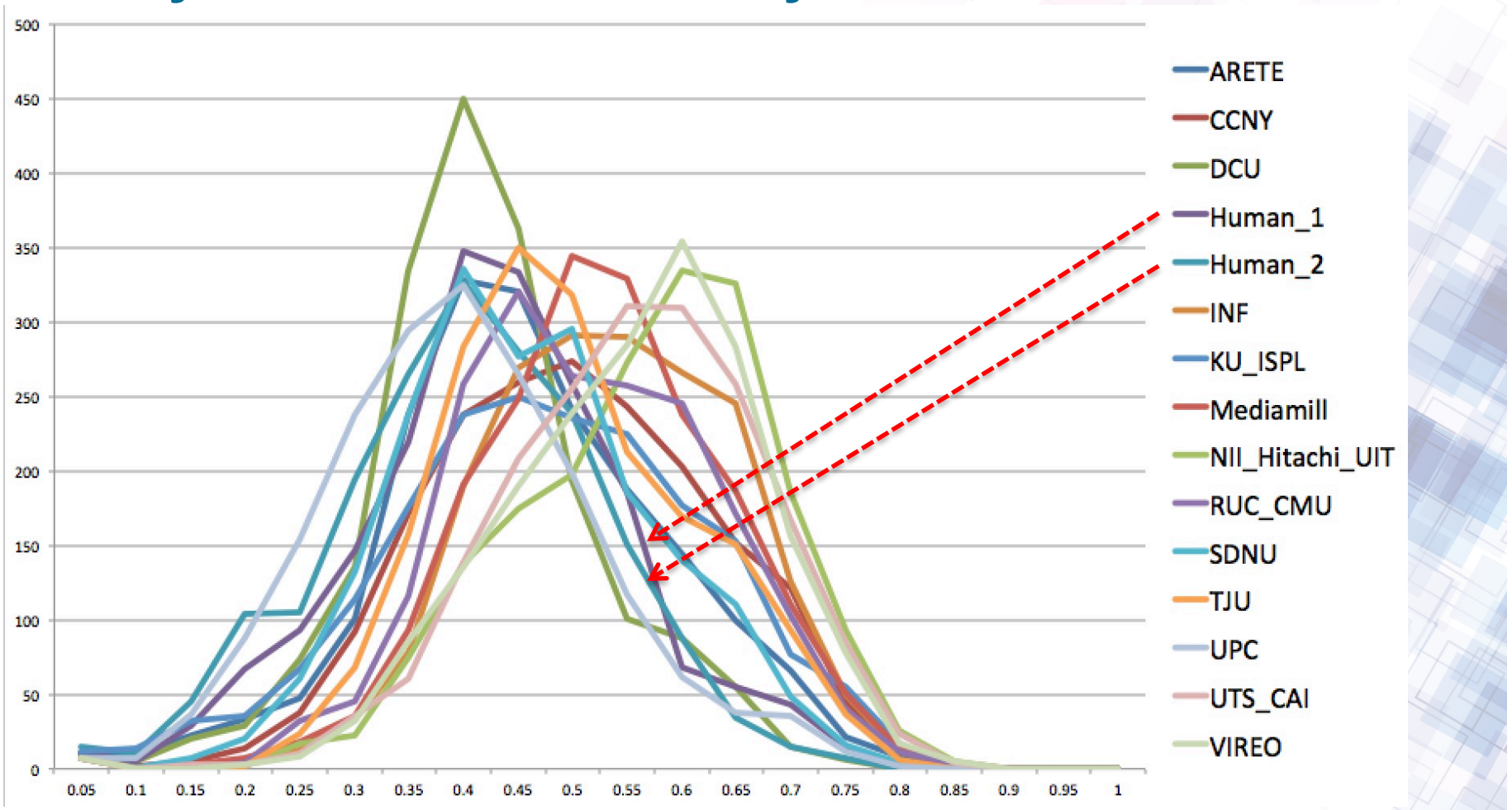
Insight
Centre for Data Analytics

Compare ARETE with, say, CCNY
… CCNY is more "with the crowd"



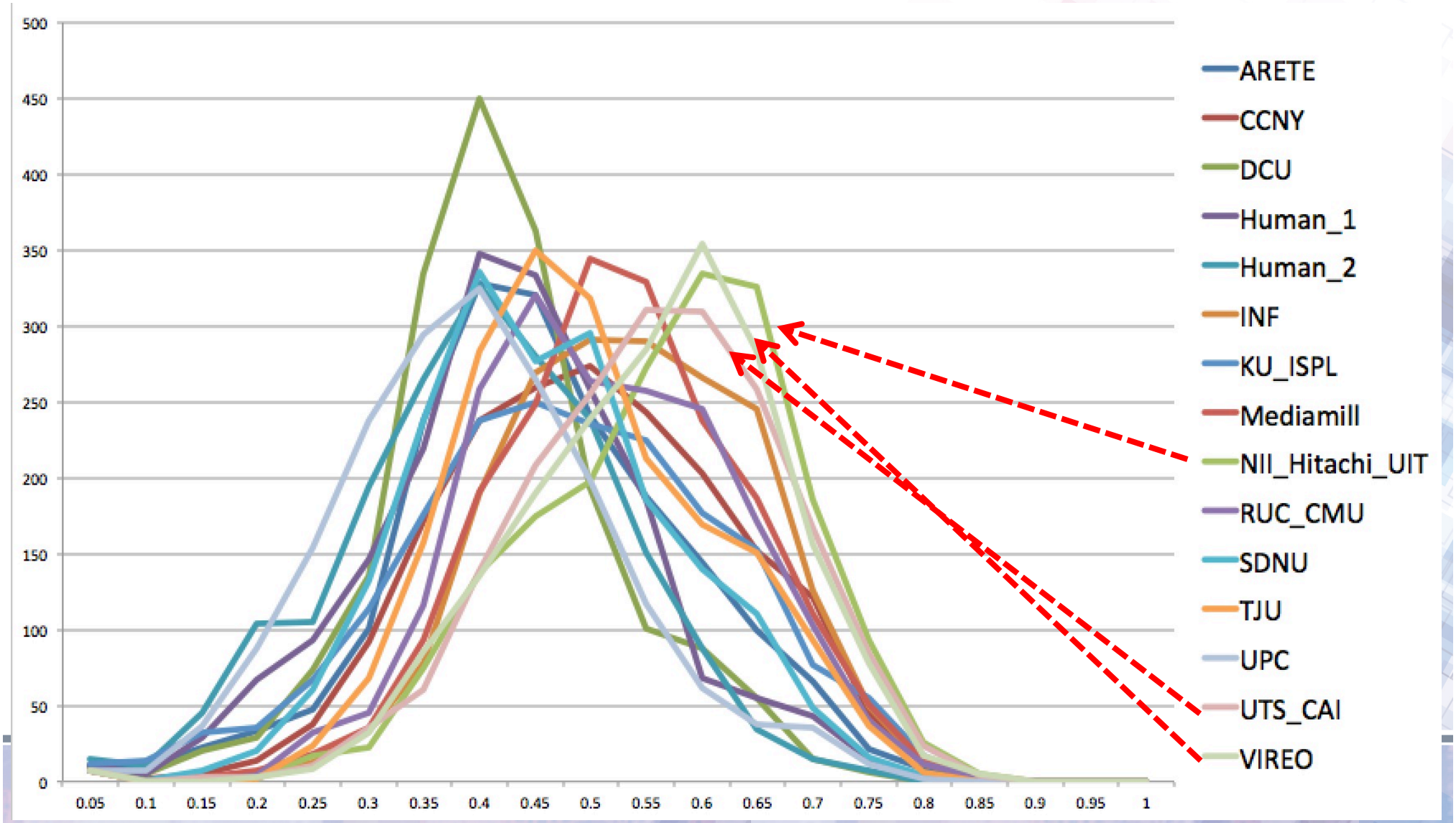
Insight
Centre for Data Analytics

Now every sumbmission vs every other, + 2x HUMAN

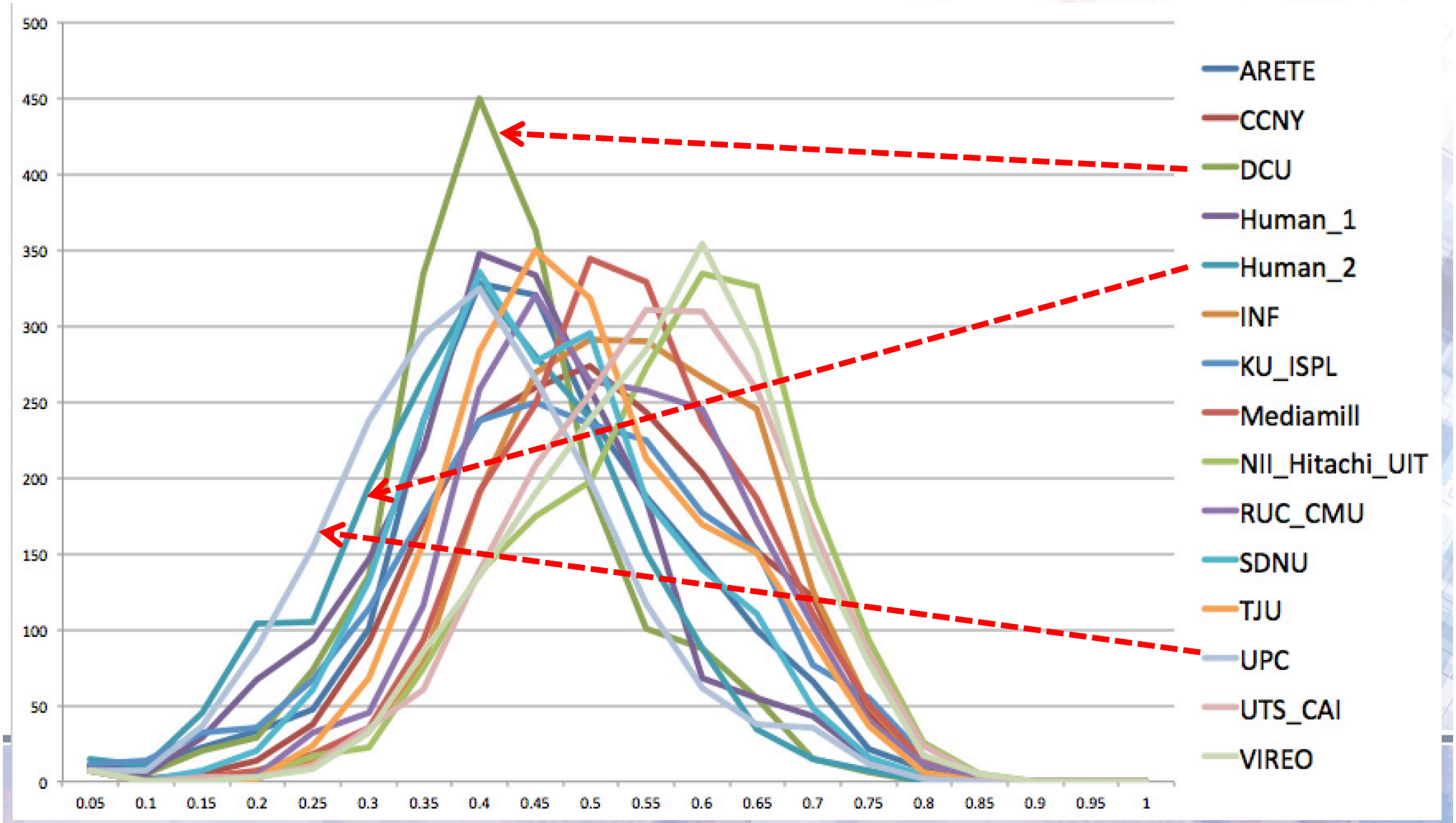


- There is an ordering – the **"popular"** systems

Insight
Centre for Data Analytics

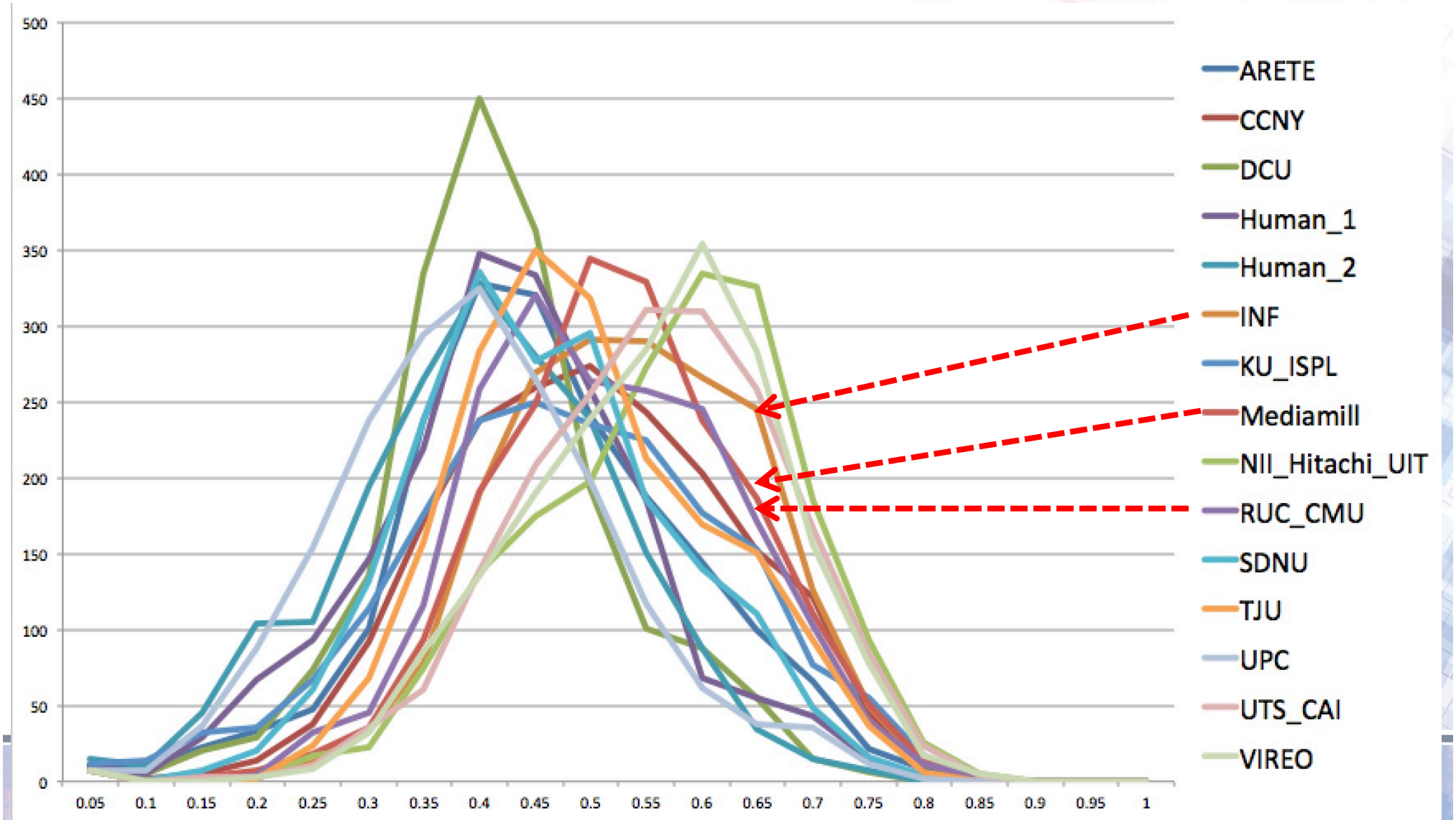


- There is an ordering – the **"outlier"** systems



- There is an ordering – the **high performing** systems

Insight
Centre for Data Analytics

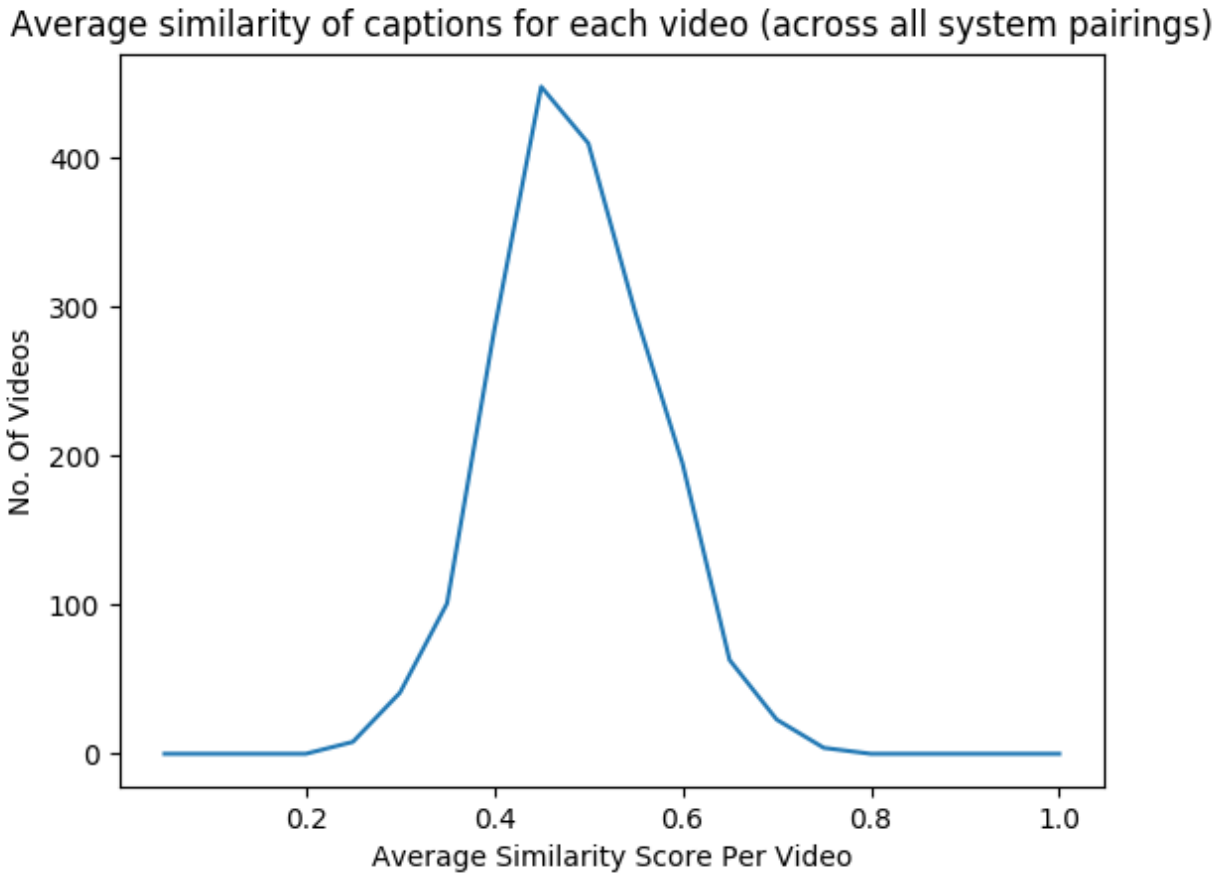


Systems Rankings for each Metric

Insight
Centre for Data Analytics

CIDEr	METEOR	BLEU	STS	DA
RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU	RUC_CMU
mediamil	mediamil	mediamil	INF	NII_Hitachi_UIT
INF	INF	TJU	mediamil	mediamil
TJU	DCU	UTS_CAI	NII_Hitachi_UIT	INF
UTS_CAI	TJU	INF	TJU	VIREO
VIREO	VIREO	DCU	UTS_CAI	UTS_CAI
NII_Hitachi_UIT	UTS_CAI	VIREO	VIREO	TJU
ARETE	KU_ISPL	NII_Hitachi_UIT	CCNY	DCU
DCU	SDNU_MMSys	SDNU_MMSys	SDNU_MMSys	CCNY
SDNU_MMSys	NII_Hitachi_UIT	CCNY	KU_ISPL	ARETE
CCNY	ARETE	ARETE	DCU	KU_ISPL
KU_ISPL	CCNY	KU_ISPL	ARETE	SDNU_MMSys
UPCer	UPCer	UPCer	UPCer	UPCer

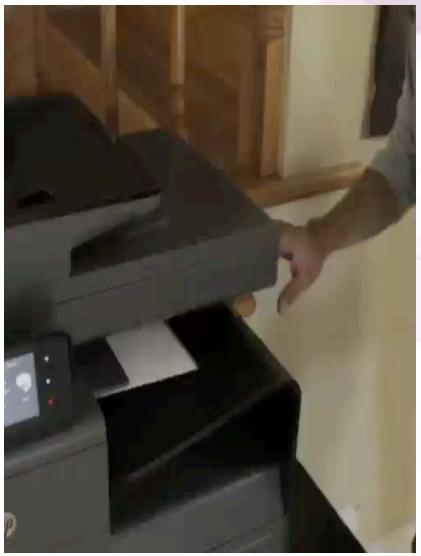
Ordering of videos by caption agreeability ?



2x Most, and least, agreed-upon videos (+ DCU captions)

1002
a woman sitting in a chair with a laptop

1457
a woman wearing a pink shirt and tie



1249
a man holding a fork and a cat

1734
a man in a suit and tie standing at a table

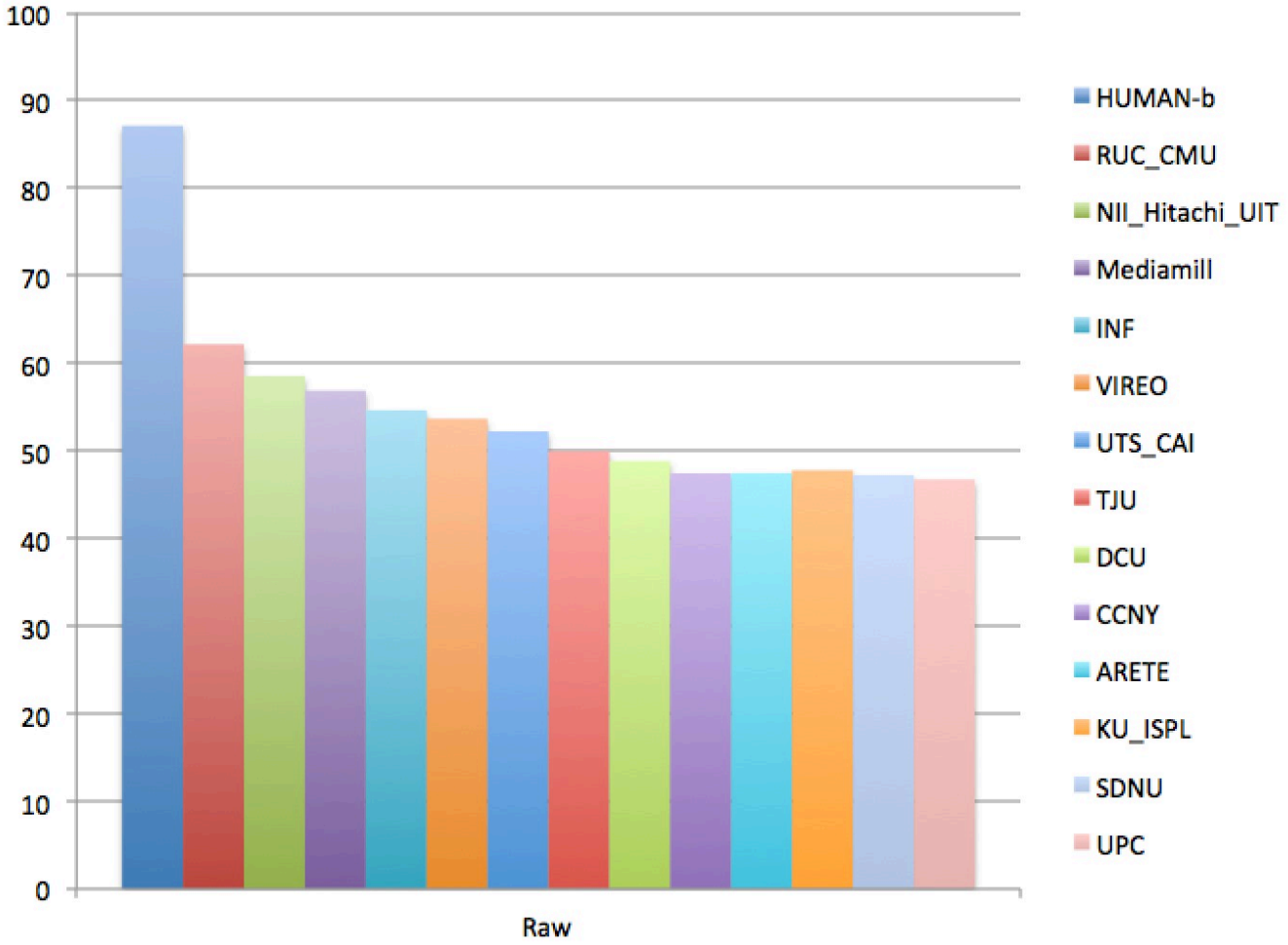
Direct Assessment Results - Analysis

- Average Direct Assessment score [0..100] for each system – micro-averaged per caption then overall average
- Also did average Direct Assessment score per system after standardisation per individual AMT worker's mean and std. deviation score, ordering unchanged

Engaging Content
Engaging People

Insight
Centre for Data Analytics

DA results - Raw



Observations

- MT metrics comparing runs against a groundtruth are flawed – DA is way to go
- Performance is good, but 25% short of ratings of captions by humans
- Approaches taken ?
    - Lots of interest in selecting **most salient parts** of videos in both spatial and temporal dimensions (we did spatial only)
    - Lots **of training data, but not enough** … MSVD (Microsoft YouTube clips), MPII-MD (Max Planck Institute), MVAD (Montreal Institute for Learning), MSR-VTT (MSR Video to Language ACM Challenge), MS-COCO (images only), TRECVid2016-VTT
    - Several (including us) used Venugopalan et al.'s ICCV 2015 Sequence to Sequence - Video to Text **(S2VT)** model
    - LSTMs and **stacked LSTMs** (us) for sentence generation
    - Several explored which is more promising for better generalization – high quality **training data or more robust models** - its the data !
    - Not many used **audio** (MFCC) segments

Future for Video Captioning …

- The metrics have changed ... STS and DA
- DA is cheap, and fast ... turnaround was 8 days from submission to graphs, and cost US$700 for all assessments
- Lots of refinement on approaches, but we can already do this quite well
- Make greater use of audio – not MFCC but Google's AudioSet
- Future tasks will include ..
    - What happens next (in a video)
    - Top-down as well as bottom up descriptions – hypothesis-driven
    - Conversational-based descriptions, beyond visual QA

Thank you.